Online Human Gesture Recognition from Motion Data Streams

Xin Zhao
School of ITEE, The University of Queensland, Australia
x.zhao@uq.edu.au

Xue Li
School of ITEE, The University of Queensland, Australia
xueli@itee.uq.edu.au

Chaoyi Pang
AEHRC, CSIRO, Australia
Chaoyi.Pang@csiro.au

Xiaofeng Zhu
College of CSIT, Guangxi Normal University, China
zhux@itee.uq.edu.au

Quan Z. Sheng
School of CS, The University of Adelaide, Australia
qsheng@cs.adelaide.edu.au

ABSTRACT

Online human gesture recognition has a wide range of applications in computer vision, especially in human-computer interaction applications. Recent introduction of cost-effective depth cameras brings on a new trend of research on body-movement gesture recognition. However, there are two major challenges: i) how to continuously recognize gestures from unsegmented streams, and ii) how to differentiate different styles of a same gesture from other types of gestures. In this paper, we solve these two problems with a new effective and efficient feature extraction method that uses a dynamic matching approach to construct a feature vector for each frame and improves sensitivity to the features of different gestures and decreases sensitivity to the features of gestures within the same class. Our comprehensive experiments on MSRC-12 Kinect Gesture and MSR-Action3D datasets have demonstrated a superior performance than the stat-of-the-art approaches.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis - *Motion*; I.5.5 [**Pattern Recognition**]: Implementation - *Interactive systems*

Keywords

Gesture Recognition, Feature Extraction, Depth Camera

1. INTRODUCTION

Human body gesture recognition has many valuable applications in computer vision such as human-computer interaction, electronic entertainment, video surveillance, patient monitoring, nursing homes, smart homes etc. The early work done by Johansson [12] suggests that the movement of the human skeleton is sufficient to be used for distinguishing different human gestures. The recent introduction of cost-effective depth camera and the related motion capturing technique [27] enables the estimation of 3D joint positions of the human skeleton, which can further generate body motion data. This phenomenon has brought on a new trend of research on body-movement gesture recognition [6, 8, 16, 32].

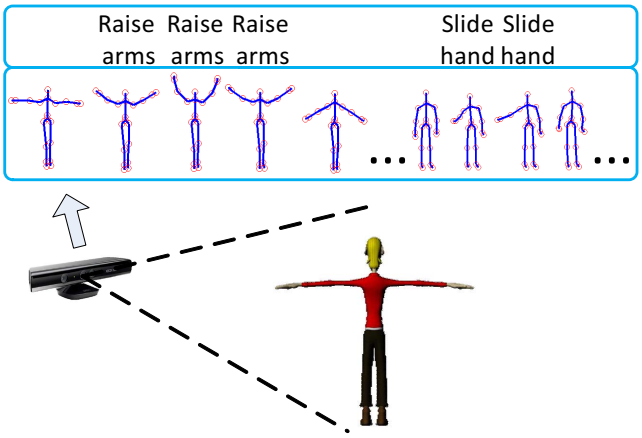


Figure 1: Scenario of online human gesture recognition from motion data stream. Depth camera is used to capture the human skeleton data stream. The gesture labels are assigned automatically to each frame by recognition. The frames without gesture label are considered as not belong to any pre-defined gesture.

Similarities exist among three concepts: *gesture*, *action*, and *activity*. The boundaries between them are not very clear. In this paper, we give our definitions based on [1].

- Gestures are elementary movements of human body parts, and are the atomic components used to describe meaningful motions of human body. One example of a gesture could be "stretching arms" or "raising legs".
- Actions are single-person activities that may be composed of multiple gestures organized temporally, such as "walking" and "waving".

MM'13, October 21–25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2502081.2502103.

- Activities refer to the interactions where two or more persons with/without objects involved.

As atomic components, gestures are less complex than actions. Research on gesture recognition lays foundation for action recognition. In this paper, we do not consider activity recognition, because objects that could be involved in activities may not be represented with the human skeleton. A scenario of online human gesture recognition from motion data stream is illustrated in Figure 1.

Gesture recognition needs to assign labels to gesture instances. Different gesture instances should be assigned with different labels, while the same gesture instances should be assigned with the same label. However, variations may occur to a gesture that has different appearances or different styles. We call this situation as *intra-class variations*. Recently, Veeraraghavan *et al.* [31] considered three sources of intra-class variations which may affect the performance of gesture recognition, namely *viewpoint*, *anthropometry*, and *execution rate*. Viewpoint variation describes the relationship between human body and the viewpoint of a camera. Anthropometry variation is related to the differences between human body sizes, and is about human physical attributes and does not change with human movements. Execution rate variation is related to temporal variability caused by the speed of human movement or by different camera frame-rates. Besides these three variations, we also advocate that the *personal style of gestures* is the fourth notable intra-class variation, which should also be considered, since different people may perform the same gesture differently.

There are two major challenges in dealing with intra-class variations. The first challenge is about how to continuously recognize gestures from unsegmented streams. Currently, most methods choose to segment gesture instances from streaming data before the recognition of gestures [8, 9]. Unfortunately, those methods suffer from the decision on the size of the segment when dealing with streaming data. By using either fixed-size or dynamic size of segments, the segmentation process itself on the streaming data introduces a new avenue of errors due to execution rate variation.

The second challenge is about the ability to differentiate intra-class variations from inter-class variations, because we need to decide whether those differences of gestures are within the same class or are between different classes. Misclassification errors may occur if we do not ignore the differences of intra-class variations or not discern the differences between classes. Online gesture recognition from motion data stream can be regarded as a problem of subsequence matching with multi-dimensional time series, where each dimension represents a specific human-body-part movement. Müller *et al.* and Sakurai *et al.* in [20, 25] proposed approaches to segment motion data stream and recognize gestures by comparing stream data with some pre-learned motion templates. A template in their approaches is at a gesture level. A template is a generic gesture instance used to match with the data stream for a class of gestures. The gesture-level motion template approaches have a major weakness on dealing with intra-class variations. Since the same gesture may have different instances because of intra-class variations, the problem of their approaches is in twofold: firstly, intra-class variations cannot be differentiated if one gesture class is represented by only one single motion template. Secondly, if every variation is to be represented by a different motion template, there must be a large number of motion templates for a single gesture class. In practice this is inefficient for dealing with real-time data streams.

A gesture may involve multiple human body parts. So a gesture is regarded as a combination of movements of human body parts. A movement is regarded as an elementary motion of one part of human body. So the granularity of motion template should be fine-tuned to be at a human-body-part movement level to reduce the redundant representation in the template modulation process. Once motion templates are represented at a human-body-part movement level, different gestures therefore, can be represented by different combinations of motion templates. So the motion templates with fine-tuned granularity can improve the efficiency of the gesture recognition.

Based on the above discussions, we consider to have a novel representation for extracting features of the human skeletons at a human-body-part movement level. This feature should also be able to represent inherent human motion characteristics such that the explicit prior segmentation process would be avoided. We call this new feature as *Structured Streaming Skeletons* (SSS). In this way, the structure of streaming skeletons is represented by a combination of human-body-part movements. So an SSS can be denoted by a vector of cardinal values of attributes that are used to describe the skeleton in a frame. Each attribute in SSS is defined as a similarity distance between the current skeleton stream and a pre-learned movement.

The two challenges involving the four intra-class variation problems mentioned above can then be dealt with by using the proposed new SSS feature as follows.

- **Viewpoint and anthropometry variations.** Motion data is generated as normalized pairwise distances of human body joints. Pairwise joints are regarded as one part of human body in this paper. Then the distances are normalized by the human body size. Therefore, SSS feature is *viewpoint* invariant and *anthropometry* invariant.

- **Execution rate variation.** The execution rate variation problem is solved by using SSS features because at each frame, each attribute is defined as the distance between the best subsequence ending at the current frame and a movement. The best subsequence is the one which is mostly similar to this movement among all subsequences ending at current frame. Different from prior segment approaches [8, 9], the size of segment can be optimized automatically during feature extraction. Therefore SSS feature is *execution rate* invariant.

- **Personal style variation.** To deal with this problem, we use *motion templates* at a granularity of human-body-part movements level. Each motion template is constructed by a human-body-part movement. Different from the approaches treating motion template at a gesture level [20, 25], we treat a template as a single dimension human-body-part movement. Therefore a gesture consists of multiple single-dimensional templates. One advantage is that different personal styles of gestures can be represented by different combinations of human-body-part movements. Therefore, SSS feature can be used to achieve *personal style* invariant.

The rest of this paper is organized as follows. Section 2 gives an overview of the related work. Section 3 describes our approach in details. Section 4 describes the experiments and the evaluations. Finally, the conclusion is given in Section 5.

2. RELATED WORK

Recognition of human gestures, actions, and activities has been extensively surveyed in recent publications [1, 5, 23, 30]. Most existing approaches [10, 14, 33, 35, 37] are about gesture and/or action recognition from color videos based on visual features, rather than the features of motion data that describe human-body-part movements. Because human-body-part movements can lead to better recognition of human gestures as well as actions, some researchers developed approaches to detect human-body-part locations first, then recognize human gestures and/or actions later. In most cases, their considerations are to recognize hand gestures online [2, 28] or recognize pre-segmented gesture instances off-line [29].

In research of online gesture and action recognition from motion data streams, Fothergill *et al.* in [8] adopted fixed-size sliding window and random forest classifiers to achieve online gesture recognition. Unfortunately, their approaches cannot handle execution rate variation and incorrect segmentation problems properly. In addition to fixed-size sliding window techniques, some researchers work on action segmentation from streaming data, then recognize the segmented action instances. Zhou *et al.* in [38] proposed a clustering algorithm to cut stream into action instances. Similarly, Gong *et al.* in [9] proposed an alignment algorithm for action segmentation. However, their approaches are all based on structure similarity between frames and are only suitable for segmenting cyclic actions. Since the structural similarity between frames of non-cyclic gestures are not always obvious, incorrect segmentation errors may occur. Incorrect segmentation will consequently introduce errors in classification process. Schwarz *et al.* in [26] generated the manifold embedding from joint positions of one frame into actions. However, the motion information is not fully considered, which may limit the approach to be scaled to more complex actions.

Template-based methods treat gesture and action recognition as a database query problem which matches data with templates in the database. Veeraraghavan *et al.* in [31] learned an average sequence and related function space of *Dynamic Time Warping* (DTW) [4] to represent each class of action. Müller *et al.* in [20] presented a procedure, where the unknown motion data is segmented and recognized by locally comparing it with available templates. The motion templates just keep the patterns of actions in the same class, with the variations ignored. Ellis *et al.* in [6] explored the trade-off between action recognition accuracy and latency. They determine key frames from motion data sequence to derive action templates. Sakurai *et al.* in [25] proposed an efficient approach to monitor streams, and to detect subsequences that are similar to a given template sequence. However, in these approaches, one action class is represented by only one template, which is insufficient to deal with intra-class variations.

Recently, Wang *et al.* in [32] proposed to learn one subset of human body joints for each action class. The subset joints are representative of one action compared to others. Additionally, they claimed that the relative positions of joints can result in more discriminative features. However, their approach is only applicable to the recognition of pre-segmented instances and cannot be used in online recognition of unsegmented data streams.

Li *et al.* in [7] proposed a *Bag of Visual Words* (BoVW) model, which is used by many researchers for action recognition from color videos, such as [24, 33]. In this paper our proposed SSS feature extraction method is different from (BoVW) model. Our SSS feature extraction method focuses on online gesture recognition from motion data stream. Each attribute of SSS feature vector has specifically defined similarity to a movement and is particularly well-suited for analyzing time-series. BoVW model uses histograms as the features for recognition of gestures. In order to count histograms, frames must be segmented first - this makes BoVW model cannot handle online gesture recognition well. More detailed discussions on the advantages of our SSS feature extraction method are given in next section.

3. PROPOSED APPROACH

Figure 2 shows the framework of our approach, which consists of two main stages: the *learning* and *prediction* stages. The goal of the learning stage is to construct a dictionary of templates and a gesture model. In prediction stage, motion data stream with unknown gestures are assigned with labels to each frame with the help of pre-learned template dictionary and the gesture model. Basically, the learning stage is off-line and prediction stage is online. We briefly describe these two stages as follows.

At learning stage, there are four steps:

1. **Motion data generation.** A training dataset captured by depth camera consists of 3D joint positions of the human skeletons. The training dataset has all gestures manually labelled on all frames. The training dataset is then scanned once. The output of this scanning is the motion data stream. The motion data stream is expressed as sequences of normalized numeric distance values of pairwise joints, which are viewpoint and anthropometry invariant. The motion data stream can be regarded as multi-dimensional time series. Each dimension represents a pair of specific human body joints. The dimensionality of motion data is determined by the number of joints that motion-capture software of depth camera can detect.

2. **Template dictionary learning.** This step is to create a dictionary of templates as a database of subsequences. We manually segment the training stream into gesture instances. Then we apply clustering algorithm to group gesture instances into a dictionary of motion templates represented as a set of subsequences. Here a template is defined as a one-dimensional time series representing distance values of two joints of human body during the time of a gesture instance. For example, in Figure 3, the motion data sequence is one instance of "slide hand" gesture. The normalized distance sequence between the joints of two hands can be a single template. As the consequence of clustering, all templates are elementary in the dictionary. We cluster each dimension of instances separately because they represent movements of different human body

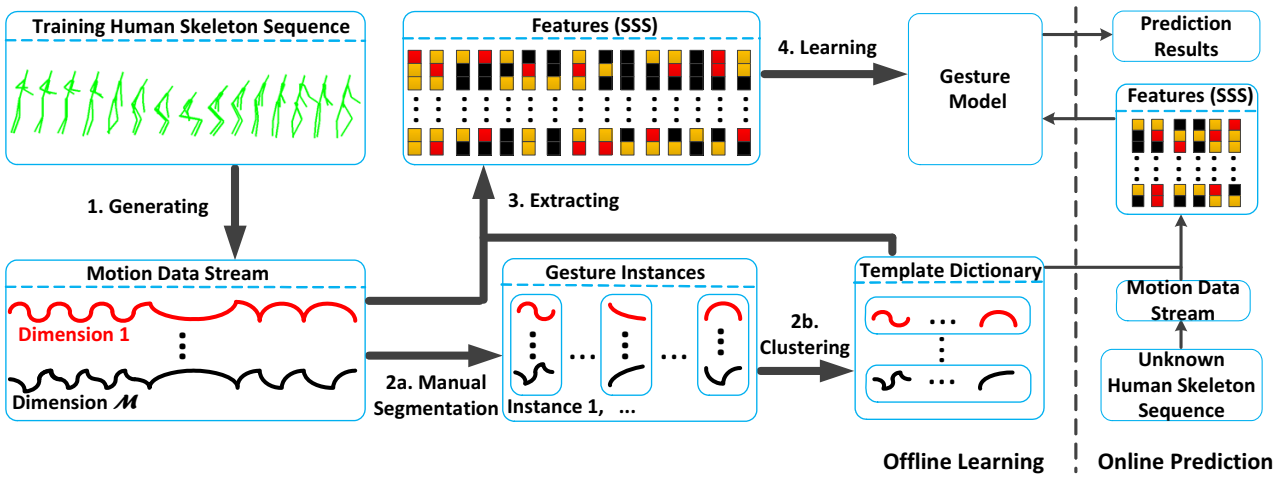


Figure 2: Framework of our approach in both learning stage and prediction stage.

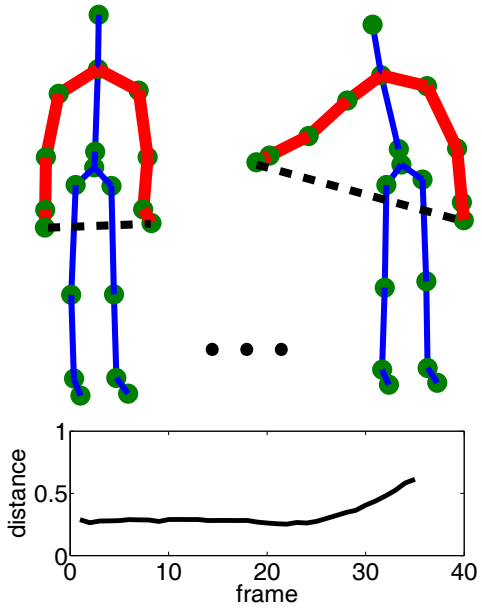


Figure 3: Example of normalized distances. The pairwise joints are "left hand" and "right hand". The upper part illustrates the skeleton sequence of one gesture instance. The lower part shows the normalized distances form a time series.

part. For ordinary human gestures, different human body parts have different movements. For example, the movement patterns of two feet may not be matched with the movements of two hands. So we cluster the movements of each human body part separately, the centroids of small number of clusters are enough to approximately represent all types of movements of this human body part. If we cluster movements of all human body parts together, like BoVW model, the required number of clusters will be very large, which decreases the efficiency of online gesture recognition. So our method has advantages over BoVM model. In our method, a gesture will be represented by a combination of a number of templates in the dictionary. This would increase the possibility to use these elementary templates to compose a large number of gestures and reduce redundancy for storage, therefore improve the efficiency of online processing.

3. **SSS feature extraction.** Based on the first two steps, a dictionary of templates is created. Now the training dataset will be scanned again for SSS feature extraction. The purpose of this step is to convert each frame into one SSS feature vector. Semantically, an SSS feature vector encodes the motion information in so-far scanned frames for the current frame. Here motion information is represented as pre-learned templates. Each SSS feature vector consists of a number of attributes represented as distance values. Each value is a minimum DTW distance between all the scanned subsequences (ending at the current frame) and a template in the dictionary for the given pair of joints. It should be pointed out that a template can only be applied to the dimension it belongs to. For example, if one template is about the joints of two hands, this template can be used to match frame sequences only on the dimension about the joints of two hands. Dimensionality of an SSS feature vector is determined by the number of templates in the dictionary. We use Figure 4 to illustrate details of SSS feature extraction that is also appeared in Figure 2.

   As indicated by Papapetrou *et al.* in [22]: if two sequences are similar to each other, their distances to template sequence are likely to be closer to each other. Similarly, if two sequences are not similar to each other, their distances to template sequence are likely to be farther from each other. Therefore, a distance value in our SSS feature is more meaningful and discriminative than histogram feature in BoVW model. SSS is specifically well-suited to the analysis of time series.

   At the end of this step the labels of gestures originally assigned to the training data frames by human become SSS feature vectors assigned to each frame and ready to be used to learn the gesture model.

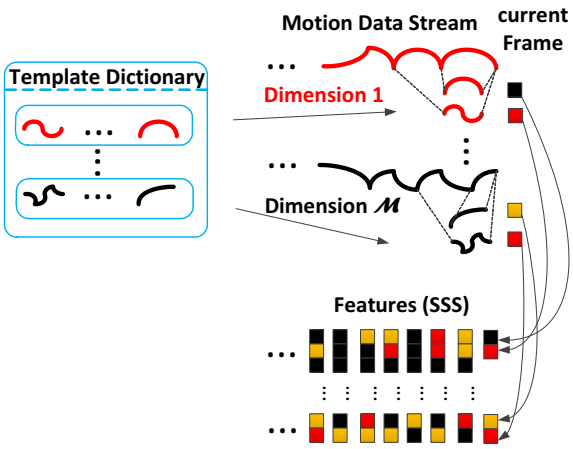


Figure 4: Example of SSS feature extraction.

4. **Gesture model training.** This step is to use a classifier to learn a gesture model from the extracted SSS feature vectors. In order to demonstrate the advantages of our proposed SSS feature extraction method, we use a basic classifier in the gesture model training. We choose a linear regression based classifier namely *Jointly Sparse Coding* [34] in this step. We will show that even using a simple classifier, the performance will be significantly improved compared with the state-of-the-art approaches.

   At the end of this step a gesture model is represented as a transformation matrix ready to be used for the prediction.

At prediction stage as illustrated in Figure 2, there are three steps performed in online prediction. Firstly, the input human skeletons captured from depth camera are translated into motion data stream using the method described in the learning stage. Then, each frame of motion data stream is mapped into an SSS feature vector also using the method described in the learning stage. Finally, at each frame, prediction is performed by a linear regression method that assigns each feature vector with a gesture label based on the learned gesture model.

3.1 Normalized Distances of Pairwise Joints

Certain amount $\mathcal{J}$ of joint positions are estimated by the motion capturing technique from depth videos. Each joint i has 3 coordinates $p_i(t) = (x_i(t), y_i(t), z_i(t))$ at frame t. For each pairwise joints i and j, $1 <= i < j <= \mathcal{J}$, we calculate their normalized distances: $s_{ij} = ||p_i - p_j||_2 / path_{ij}$, where $path_{ij}$ is the path between joints i and j in the human skeleton. As shown in the upper part of Figure 3, $\mathcal{J} = 20$ joints are captured with Microsoft Kinect system [27], and $\mathcal{M} = \mathcal{J} \times (\mathcal{J}-1)/2 = 190$ pairwise joints are generated. We take "left hand" and "right hand" for example. The dotted line is the Euclidean distance between them. The bold lines indicate the path of these two joints in the human skeleton. We can see that s_{ij} has no relationship with the body position, body orientation and body size. *i.e.*, viewpoint and anthropometry invariant.

We treat the normalized distance of one pairwise joints as one dimension of motion data. Along the time axis, the distances form a one-dimensional time series, as shown in the lower part of Figure 3. Therefore, motion data stream is a multiple dimensional time series: $\mathbf{S}(:, t) = \{s_{ij}(t)\}, 1 <= i < j <= \mathcal{J}$.

3.2 Template Dictionary

From training motion data stream $\mathbf{S} = [\mathbf{S}(:,1), ..., \mathbf{S}(:,\mathcal{N})]$, where $\mathcal{N}$ is the number of frames in motion data stream $\mathbf{S}$, we manually segment all gesture instances and learn a template dictionary $\mathbf{D}$. One template is one dimension time series of one instance. For each dimension of motion data, we cluster these instances into $\mathcal{G}$ clusters. In each cluster, the gesture instance with minimum average distance to others on this dimension is chosen as one template. There are $\mathcal{G} \times \mathcal{M}$ templates in this dictionary. Therefore, template dictionary $\mathbf{D} = \{d_{ij}^g\}$, where i and j indicate the pairwise joints and g indicates the cluster index in this pairwise joints. Spectral clustering algorithm [21], which can be used in non-Euclidean space, is adopted for the clustering. K-means clustering algorithm [11] is not applicable to non-Euclidean space, so it is not suitable here. We also tested other clustering algorithms such as k-medoids clustering [13] and optics clustering [3]. Spectral clustering algorithm is the most robust one. DTW is adopted as the distance measure for time series to eliminate execution rate variation.

3.3 New SSS Feature

For recognition, each frame in $\mathbf{S}$ is required to be represented by an SSS feature vector. We compute the distances between the best fitting subsequences ending at this frame and template dictionary $\mathbf{D}$ as the SSS feature vector.

We use $\mathbf{S}(:, \tilde{t} : t)$ to represent all subsequences ending at frame t. From the dictionary $\mathbf{D}$, each template d_{ij}^g is used to find one best fitting subsequence $\mathbf{S}(:, \hat{t}_{ij}^g : t)$ among $\mathbf{S}(:, \tilde{t} : t)$. The distance between the template and the best fitting subsequence on related dimension is the minimum:

$$x_{ij}^g(t) = |s_{ij}(\hat{t}_{ij}^g : t) - d_{ij}^g| \quad (1)$$

$$\hat{t}_{ij}^g = \arg\min_{\tilde{t}} |s_{ij}(\tilde{t} : t) - d_{ij}^g| \quad (2)$$

Here, DTW is still used as the distance measure between two sequences to eliminate execution rate variation. The stream monitoring technique [25] can be used to detect the optimal starting point $\hat{t}_{ij}^g$.

We further use an SSS feature vector to represent one frame. Each minimum distance x_{ij}^g according to one template d_{ij}^g is one attribute of the SSS feature vector. In this paper, $\mathcal{G} \times \mathcal{M}$ templates in the dictionary $\mathbf{D}$ can generate a vector with $\mathcal{G} \times \mathcal{M}$ dimensions. We treat this vector $\mathbf{X}(:, t) = \{x_{ij}^g(t)\}$ as the SSS feature for frame t. $\mathbf{X} = [\mathbf{X}(:,1), ..., \mathbf{X}(:,\mathcal{N})]$ is the SSS feature matrix for $\mathbf{S}$.

3.4 Classification

It is insufficient to use only one attribute of the feature to discriminate different gestures. How to learn a combination of these attributes, which is representative of identical gesture and discriminative compared with different gesture. This can be achieved by machine learning techniques. We adopt jointly sparse coding [34], which focus on feature selection and classification, to learn the combination.

Table 1: Notations.

Symbol	Description
$\mathcal{J}$	number of joints
$\mathcal{M}$	number of joint pairs
$\mathcal{N}$	number of frames
$\mathcal{G}$	number of clusters
$\mathbf{S}$	motion data stream
$\mathbf{D}$	template dictionary
$\mathbf{X}$	SSS feature matrix
$\mathbf{W}$	gesture model

Classic least square regression is used to learn a transformation matrix $\hat{\mathbf{W}}$ to transfer features into gesture labels. Furthermore, without reducing the effectiveness, the amount of involved attributes should be as small as possible to improve the efficiency. This can be achieved with $\ell_{2,1}$ norm regularization. In addition, the minimization of $\ell_{2,1}$ norm regularization will enable the algorithm to leverage the shared information across multiple gesture classes [18].

$$\hat{\mathbf{W}} = \underset{\mathbf{W}\in\mathbb{R}^{(\mathcal{G}\times\mathcal{M})\times\mathcal{C}}}{\arg\min} ||\mathbf{W}^\top\mathbf{X} - \mathbf{Y}||_F^2 + \lambda||\mathbf{W}||_{2,1} \qquad (3)$$

where $\mathcal{C}$ is the number of gesture classes, matrix $\mathbf{Y} = [\mathbf{Y}(:,1), ..., \mathbf{Y}(:,\mathcal{N})]$, $\mathbf{Y}(:,t) \in \mathbb{R}^{\mathcal{C}}$ indicates the multiple labels for frame t in stream $\mathbf{S}$. If $\mathbf{S}(:,t)$ belongs to the c^{th} gesture class, $\mathbf{Y}(c,t) = 1$, $\mathbf{Y}(e,t) = -1$, for $e \neq c$. If $\mathbf{Y}(:,t)$ does not belong to any $\mathcal{C}$ classes of gestures, $\mathbf{Y}(:,t) = -1$, matrix $\hat{\mathbf{W}}$ is the gesture model, with the help of $\ell_{2,1}$ norm regularization, many rows of $\hat{\mathbf{W}}$ are near to 0 [36]. We can prune invalid feature attributes. In this paper, the weight of the m^{th} attribute in feature is measured with $||\hat{\mathbf{W}}(m,:)||_2$ We descendingly sort these weights. The first $\mathcal{K}$ attributes who's weight sum is up to 99% of the total are regarded as valid. Others are set to 0, and related template are noted as invalid templates. Here λ is the parameter to control regularization.

3.5 Prediction

In online prediction stage, given an unknown motion data steam, at frame t, we extract SSS feature $\mathbf{U}(:,t)$ only with the valid templates in dictionary $\mathbf{D}$. The attributes related with invalid templates are set to 0 without computation. If $max(\hat{\mathbf{W}}^\top\mathbf{U}(:,t)) \geq \beta$, the row index with maximum value indicates the gesture class, otherwise, this frame does not belong to any $\mathcal{C}$ classes of gestures. Here β is a parameter for leverage precision and recall.

In online prediction stage, at each frame, the time complexity of generating motion data is $O(\mathcal{M})$, the time complexity of extracting SSS feature is $O(\mathcal{M}\times\mathcal{K}\times\mathcal{A})$, where $\mathcal{A}$ is the average length of template sequence, and the time complexity of classification is $O(\mathcal{C}\times\mathcal{K})$.

The notations used in this paper are given in Table 1.

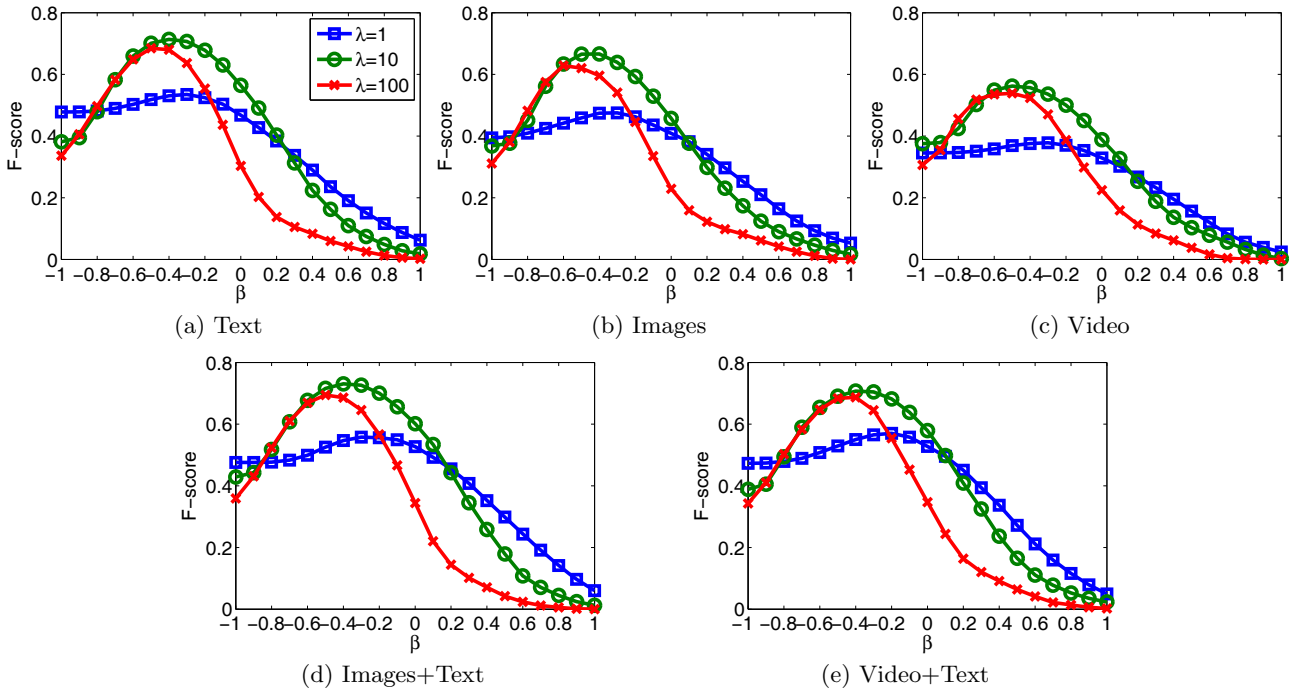
4. EXPERIMENTS

We chose MSRC-12 Kinect Gesture dataset [8] to evaluate our proposed SSS feature for online gesture recognition. To the best of our knowledge, this is the only one public dataset for the research of online human gesture recognition from motion data stream. However, this dataset is very large with more than 700,000 frames available for the ground truth testing. In addition, we also validated our approach in pre-segmented action recognition using MSR-Action3D dataset [15] which is a well-known action recognition dataset for bench-marking with relevant algorithms. However, the data has been pre-segmented for evaluating action instances already. So the advantages of our SSS feature on online prediction cannot be demonstrated. Moreover we can still use this dataset for the demonstration of the advantages of our SSS feature on the lower level granularity of recognition.

4.1 Results on MSRC-12 Kinect Gesture Dataset

MSRC-12 Gesture dataset comprises of 594 sequences, more than 700,000 frames (approximately 6 hours and 40 minutes) collected from 30 people performing 12 classes of gestures. In total, there are 6,244 gesture instances. The ending points of all gesture instances were manually labeled. Twenty human body joints ($\mathcal{J} = 20$) are captured with Microsoft Kinect system. The body poses are captured at a sample rate of 30Hz with an accuracy approximately two centimeters in joint positions. In this dataset, for research various methods of teaching human on how to perform different gestures, the participants were provided with three instruction modalities and their combinations to perform gestures. The three instruction modalities are i) text descriptions, ii) image sequences, and iii) video demos. There are also two combinations of the three modalities, *i.e.*, images with text, and video with text. When participants are given instructions, different modalities of instructions may cause different responses. For example, when instructions are given by videos, the gestures can be performed with an imitated pace from the images of video. This would cause the fixed-size sliding window approach can have less chances of segmentation errors.

For a comparison with the state-of-the-art work in [8], we use the same criteria (F-score) and latency-aware measure as that justified in [8]. Precision measures how often is the gesture actually present when the system predicts it is. Recall measures how many true gestures are recognized by the approach. Latency measures how much time is the delay between the true action starting point and the prediction. Following the experiment setting of [8] , we treat the previous 34 frames and ending point as one gesture instance. Thus, the average length of templates is $\mathcal{A} = 35$ frames. For a specified amount of tolerated latency $\triangle = 10$ frames, a fixed window of size $2\triangle$ is centered around each ending point. All the frames inside the window are given the same gesture label as the ending point, and other frames outside the window are regarded as negative samples. Therefore, the latency is 0.83s-1.5s ((35 ± 10)(frames)/30(Hz)). In this way, we obtain the ground truth label of each frame for evaluation. Each frame is treated as one sample for training and test. A balanced F-score between 0 and 1 combines precision and recall is chosen as the evaluation measure. We measure the intra-modality generalization performance: training and testing using the same instruction modality. For each modality, a "leave-person-out" protocol is used to split dataset into 10 disjoint sets. Here, disjoint is in terms of the person-gesture combination. In each set, we remove a set of persons from full dataset to obtain the minimum set that contains performances of all gestures. Nine sets are used for training, and one set is used for testing. We finally obtain five F-scores, one for each modality. Each F-score is an average over 10 repetitions and 12 gestures.



(a) Text (b) Images (c) Video

(d) Images+Text (e) Video+Text

Figure 5: The relationship between recognition performance and parameters λ and β. The optimal parameters are stabilized at $\lambda = 100$ and $-0.5 \leq \beta \leq -0.3$.

Table 2: Comparison on MSRC-12 Gesture dataset on five modalities.

Method	Text	Images	Video	Images+Text	Video+Text
Ours	$\mathbf{0.713 \pm 0.191}$	$\mathbf{0.666 \pm 0.194}$	0.557 ± 0.291	$\mathbf{0.730 \pm 0.148}$	$\mathbf{0.707 \pm 0.17}$
Baseline [8]	0.479 ± 0.104	0.549 ± 0.102	$\mathbf{0.627 \pm 0.053}$	0.563 ± 0.045	0.679 ± 0.035

We compare our approach with the state-of-the-art method [8], which adopted fixed size sliding window and random forest classifiers for this dataset. In this dataset, we use fixed values $\lambda = 100$ and $\beta = -0.4$ for evaluation. These parameters are derived from an experiment that is designed to find the optimal parameters. In that experiment, parameter λ was firstly set as $\{10, 100, 1000\}$ and parameter β was set as $\{-1, -0.9, -0.8, ..., 0.8, 0.9, 1\}$. The dataset is also split into 10 disjoint sets (different from the 10 disjoint sets split for the evaluation) for finding the optimal parameters (λ and β). The reason we do not use conventional 10-fold cross validation is that it is very time consuming when dataset is large (with more than 700,000 samples in this dataset), so the parameters can be derived by an experiment. In fact, Fothergill *et al.* in [8] also chose the parameters in this way.

We show the effect caused by the parameter λ and parameter β when number of clusters $\mathcal{G}$ is fixed to 20 (the reason to choose parameter $\mathcal{G}$ as 20 will be discussed in following). The recognition F-scores of all modalities are illustrated in Figure 5. We can see that the optimal values of parameters stabilize at $\lambda = 100$ and $-0.5 \leq \beta \leq -0.3$.

Table 2 shows the comparison results for each modality. We can see that our approach obtains average improvement of F-scores by 10%. There are considerable improvements in "Text", "Images", and "Images+Text" modalities, which are more susceptible to execution rate variation. This demonstrates that our approach can handle incorrect segmentation problems. It worths to note that the instructions given by video modality is impractical in an experiment because the variations of different participants are mostly precluded by the behaviors of participants who are simply imitating the video instructions. In human-computer interaction applications, user should remember instructions by themselves, they have no chance to imitate the gestures by watching video instructions. In this case, the high F-score of video modality has little significance.

We notice that our approach presents more variance than that of the baseline. One explanation is that parameter β causes the variance. In our experiments, the F-score is the average over 10 repetitions and 12 gestures, with the same value of parameter β. The optimal value of parameter β is stable for the highest average F-score, but for different gestures in different repetitions, the optimal value of parameter β changes. In our current system, we adopt a linear regression classifier to demonstrate the advantages of our proposed SSS feature. It works well even with a simple classifier. In our future work, we will consider the model selection problem for our proposed SSS feature.

We also notice that "Text" exhibits better performance. As Fothergill *et al.* in [8] reported "the text provided a specific of what the sensor was going to pick up". So people can more clearly understand the key movements of differ-

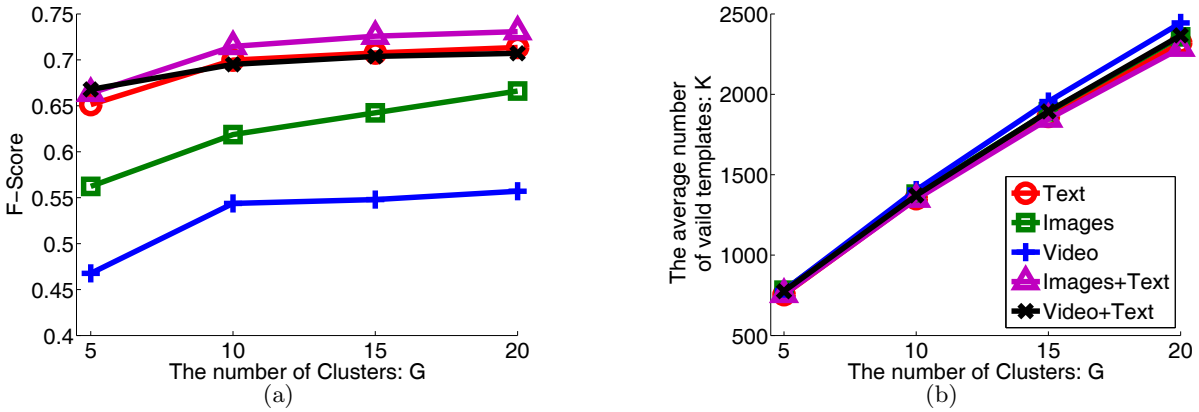


Figure 6: (a) The relationship between recognition performance and the number of clusters $\mathcal{G}$. (b) The relationship between the number of valid templates $\mathcal{K}$ and the number of clusters $\mathcal{G}$.

ent gestures. However, the research on the effect caused by modality of instruction is beyond the scope of this paper. For more details about MSRC-12 Gesture dataset, please refer to [8].

Furthermore, we research the effect caused by the number of clusters. We set $\mathcal{G} = [5, 10, 15, 20]$ respectively when $\lambda = 100$ and $\beta = -0.4$. The recognition F-scores and the number of valid templates $\mathcal{K}$ are illustrated in Figure 6. As the increasing number of clusters, F-scores increase slowly, and the number of valid templates increases linearly. Therefore, we choose $\mathcal{G} = 20$ to balance the effectiveness as well as the efficiency.

We preform our experiments with hardware of "i7 860 CPU" and "4G RAM", and software of Matlab hybrid with parts of C code. With our approach, In prediction stage, gesture recognition of one frame costs less than 2ms.

4.2 Results on MSR-Action3D Dataset

MSR-Action3D dataset [15] comprises of 557 pre-segmented action instances. There were 10 people performing 20 classes of gestures. Same as MSRC-12 Kinect Gesture Dataset, human body joints ($\mathcal{J} = 20$) were captured with Microsoft Kinect system.

Because the instances have been manually segmented, we simplify online extracting features as computing the distances between the pre-segmented instance and the template dictionary directly. Each instance is treated as one sample. We use fixed values $\lambda = 10$ and $\mathcal{G} = 20$ for experiments. In this dataset, the parameters are optimized on the test sets. In all comparing approaches, the parameters are optimized in the same way. The fairness of the comparison is evidenced by using the same experiment setting (the method of partition of training datasets and test datasets, and the method of parameter tuning) on the same standard dataset.

We compare our approach with the state-of-the-art methods on the cross-subject test setting [15, 32], where the samples of half number of persons are used as training data, and the rest are used as testing data. As Table 3 shows, our approach outperforms the other time series based methods [6, 17, 19, 20], which treat the motion data as an undivided whole set. The only approach [32] that outperforms ours uses a subset of joints for classification, which is similar to our approach, but it focuses on recognition at pre-segmented document level, and cannot be used in online recognition from unsegmented streams.

Table 3: Comparison on MSR-Action3D dataset.

Method	Accuracy
Recurrent Neural Network [19]	0.425
Dynamic Temporal Warping [20]	0.54
Hidden Markov Model [17]	0.63
Multiple Instance Learning [6]	0.657
Our Approach	**0.817**
Actionlet Ensemble [32]	0.882

The confusion matrix is illustrated in Figure 7. We can see that for most actions, our approach works well, while for the similar actions such as "hand catch" and "high throw", "draw X" and "draw circle", there are some misclassifications. It can be seen that, for each action, there are about 10 instances performed by 5 people for training, which may be insufficient to distinguish these similar gestures.

5. CONCLUSIONS

Depth cameras are now widely used in applications of human-computer interaction. There is a growing need to apply depth cameras in human behaviors detections, such as gesture, action, and activity recognition. The effective and efficient recognition of human gestures in a real-time fashion has a significant impact on the recognition of human actions.

In a nutshell, our **contributions** are as follows:

- **New SSS feature.** We proposed a novel feature, namely, *Structured Streaming Skeletons* (SSS), for online gesture recognition from motion data streams to deal with four types of intra-class variations (*i.e.*, viewpoint, anthropometry, execution rate, and personal style), thereby effectively and efficiently solved the incorrect segmentation and inadequate template matching problems.

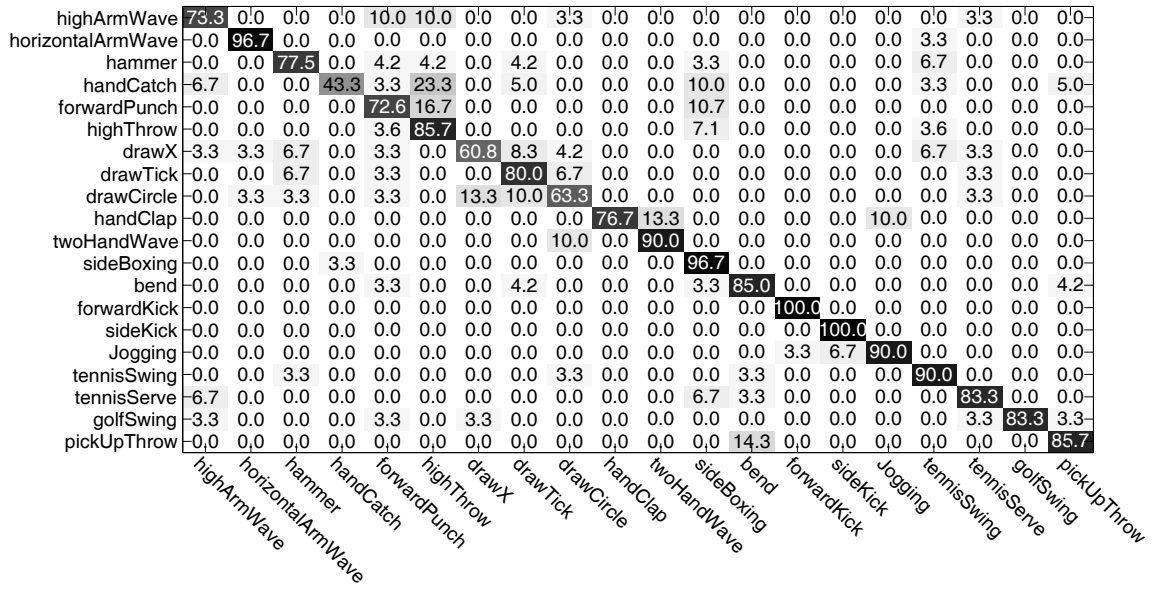


Figure 7: The confusion matrix of our proposed approach for MSR-Action3D dataset.

- **None prior segmentation.** We detect the size of segment by dynamically matching with pre-learned templates. Execution variation is eliminated and there is no avenue for errors made by prior segmentations.
- **Fine-tuned granularity of motion templates.** We create a motion template dictionary at a granularity of elementary body-part-movement level. We consider human body as a combination of many small parts and perform body part analysis separately. One advantage is that personal styles of gestures can be represented by different combinations of human-body-part movements.
- **Superior online performance.** Because of the discriminative nature of SSS feature, the superior performance is achieved even with a simple classifier, with average improvement of F-scores by 10% (Table 2) compared with the stat-of-the-art approaches. Also our online prediction of gestures is extremely fast costing less than 2ms per frame. The latency is 0.83s-1.5s (in realtime response)[1].

Our further research will consider: i) the model selection based on our proposed SSS feature; ii) online gesture recognition with inaccurate skeleton data to reduce gesture recognition errors that are caused by incomplete skeleton tracking; iii) studies of real user experience.

6. ACKNOWLEDGEMENTS

This work was partially supported by the Australian Research Council (ARC) Discovery Project DP130104614.

7. REFERENCES

[1] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, 43(3):16, 2011.

[2] J. Alon, V. Athitsos, Q. Yuan, and S. Sclaroff. A unified framework for gesture recognition and spatiotemporal gesture segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 31(9):1685–1699, 2009.

[3] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. Optics: ordering points to identify the clustering structure. *ACM SIGMOD International Conference on Management of Data (SIGMOD)*, 28(2):49–60, 1999.

[4] D. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370, 1994.

[5] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding (CVIU)*, 117(6):633–659, 2013.

[6] C. Ellis, S. Masood, M. Tappen, J. LaViola, and R. Sukthankar. Exploring the trade-off between accuracy and observational latency in action recognition. *International Journal of Computer Vision (IJCV)*, 101(3):420–436, 2013.

[7] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 524–531, 2005.

[8] S. Fothergill, H. M. Mentis, P. Kohli, and S. Nowozin. Instructing people for training gestural interactive systems. In *ACM annual conference on Human Factors in Computing Systems (CHI)*, pages 1737–1746, 2012.

[9] D. Gong, G. Medioni, S. Zhu, and X. Zhao. Kernelized temporal cut for online temporal segmentation and recognition. In *European Conference on Computer Vision (ECCV)*, pages 229–243, 2012.

[1]Our system demo video is attached as the supplementary material on YOUTUBE: http://youtu.be/l4zzmrXfdag

[10] T. Guha and R. K. Ward. Learning sparse representations for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(8):1576–1588, 2012.

[11] J. Hartigan and M. Wong. A k-means clustering algorithm. *Journal of the Royal Statistical Society C*, 28:100–108, 1979.

[12] G. Johansson. Visual motion perception. *Scientific American*, 232(6):76–88, 1975.

[13] L. Kaufman and P. Rousseeuw. Clustering by means of medoids. *Statistical data analysis based on the L1-norm and related methods*, pages 405–416, 1987.

[14] H. Li and M. Greenspan. Model-based segmentation and recognition of dynamic gestures in continuous video streams. *Pattern Recognition*, 44(8):1614–1628, 2011.

[15] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *CVPR Workshop*, pages 9–14, 2010.

[16] S.-Y. Lin, C.-K. Shie, S.-C. Chen, and Y.-P. Hung. Action recognition for human-marionette interaction. In *ACM international conference on Multimedia (MM)*, pages 39–48, 2012.

[17] F. Lv and R. Nevatia. Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In *European Conference on Computer Vision (ECCV)*, pages 359–372, 2006.

[18] Z. Ma, Y. Yang, Y. Cai, N. Sebe, and A. G. Hauptmann. Knowledge adaptation for ad hoc multimedia event detection with few exemplars. In *ACM international conference on Multimedia (MM)*, pages 469–478, 2012.

[19] J. Martens and I. Sutskever. Learning recurrent neural networks with hessian-free optimization. In *International Conference on Machine Learning (ICML)*, pages 1033–1040, 2011.

[20] M. Müller, A. Baak, and H.-P. Seidel. Efficient and robust annotation of motion capture data. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA)*, pages 17–26, 2009.

[21] A. Ng, M. Jordan, Y. Weiss, et al. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems (NIPS)*, pages 849–856, 2002.

[22] P. Papapetrou, V. Athitsos, M. Potamias, G. Kollios, and D. Gunopulos. Embedding-based subsequence matching in time-series databases. *ACM Transactions on Database Systems (TODS)*, 36(3):17, 2011.

[23] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, 2010.

[24] M. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1036–1043, 2011.

[25] Y. Sakurai, C. Faloutsos, and M. Yamamuro. Stream monitoring under the time warping distance. In *IEEE International Conference on Data Engineering (ICDE)*, pages 1046–1055, 2007.

[26] L. Schwarz, D. Mateus, V. Castañeda, and N. Navab. Manifold learning for tof-based human body tracking and activity recognition. In *British Machine Vision Conference (BMVC)*, pages 1–11, 2010.

[27] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1297–1304, 2011.

[28] Y. Song, D. Demirdjian, and R. Davis. Continuous body and hand gesture recognition for natural human-computer interaction. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(1):5, 2012.

[29] K. Tran, I. Kakadiaris, and S. Shah. Part-based motion descriptor image for human action recognition. *Pattern Recognition*, 45(7):2562–2572, 2012.

[30] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 18(11):1473–1488, 2008.

[31] A. Veeraraghavan, R. Chellappa, and A. K. Roy-Chowdhury. The function space of an activity. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 959–968, 2006.

[32] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1297, 2012.

[33] S. Wang, Y. Yang, Z. Ma, X. Li, C. Pang, and A. G. Hauptmann. Action recognition by exploring data distribution and feature correlation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1370–1377, 2012.

[34] Y. Yang, Z. Ma, A. G. Hauptmann, and N. Sebe. Feature selection for multimedia analysis by sharing information among multiple tasks. *IEEE Transactions on Multimedia (TMM)*, 15(3):661–669, 2013.

[35] Y. Yang, I. Saleemi, and M. Shah. Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(7):1635–1648, 2013.

[36] Y. Yang, Y. Yang, Z. Huang, H. Shen, and F. Nie. Tag localization with spatial correlations and joint group sparsity. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 881–888, 2011.

[37] Z. Zhang and D. Tao. Slow feature analysis for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(3):436–450, 2012.

[38] F. Zhou, F. Torre, and J. Hodgins. Aligned cluster analysis for temporal segmentation of human motion. In *IEEE Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–7, 2008.